DOCUMENT RESUME

ED 442 836 TM 031 256

AUTHOR Kobrin, Jennifer L.
TITLE An Investigation of the Cognitive Equivalence of Computerized and Paper-and-Pencil Reading Comprehension Test Items.
PUB DATE 2000-04-26
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Cognitive Processes; *College Students; *Computer Assisted Testing; Higher Education; *Protocol Analysis; *Reading Comprehension; Reading Tests; *Student Attitudes; *Test Format; Test Results
IDENTIFIERS Graduate Record Examinations; *Paper and Pencil Tests

ABSTRACT

The comparability of computerized and paper-and-pencil tests was examined from cognitive perspective, using verbal protocols rather than psychometric methods, as the primary mode of inquiry. Reading comprehension items from the Graduate Record Examinations were completed by 48 college juniors and seniors, half of whom took the computerized test first followed by the paper-and-pencil version, and half of whom took the paper-and-pencil test before the computerized test. Participants were asked to think aloud as they answered the test questions. The verbal protocols were transcribed and coded for interpretation. There was a greater frequency of reading comprehension utterances during the paper-and-pencil test, but these were largely accounted for by the use of physical aids to identify important information in the passage. Many participants said that they felt disadvantaged during the computerized test by not being able to write on the passage and test questions. The frequently used strategy of marking the test did not seem to produce any cognitive benefits, however. There was slight evidence of a working memory load while answering the questions on the computerized tests, but overall there were few mode differences and the magnitude of differences was very small. Nearly all participants used the same overall test-taking strategy on both test formats. The first test given, which was less interesting and more difficult, exposed more of the mode effects than the more interesting second test. An appendix contains a chart of coding categories at the utterance level. (Contains 10 tables and 30 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 442 836

Running head: COGNITIVE EQUIVALENCE OF TEST ITEMS

An Investigation of the Cognitive Equivalence of Computerized and Paper-and-Pencil Reading Comprehension Test Items

Jennifer L. Kobrin

Rutgers University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

J. Kobrin

1

Paper presented at the Annual Meeting of the American Educational Research Association

New Orleans, Louisiana

April 26, 2000

TM031256

BEST COPY AVAILABLE

An Investigation of the Cognitive Equivalence of Computerized and Paper-and-Pencil Reading Comprehension Test Items

Introduction

Computers have become an integral part of education, as their use has impacted on nearly every aspect of instruction. The increased availability and capacity of computing resources has also led to a revolution in educational measurement, with the popular use of computers to construct, deliver, and score educational and psychological tests. Although there are many benefits associated with delivering tests on the computer, there is the potential that the mode of delivery changes the constructs that the test was designed to measure. The methods employed to determine equivalence have consisted predominantly of correlational studies and the comparison of mean total scores and/or individual item scores obtained by examinees on parallel computerized and paper-and-pencil versions of the same test. For the most part, these studies have reported that the scores on the two test modes are very similar and that the correlations are moderate to high (Mazzeo & Harvey, 1988; Mead & Drasgow, 1993). However, establishing that examinees receive comparable scores on parallel computerized and paper-and-pencil tests is not enough to verify that the two test modes have equal construct validity.

An important part of the test validation process is determining whether a test includes construct-irrelevant test variance, that is "excess reliable variance that is irrelevant to the interpreted construct (Messick, 1989, p. 34)." Construct-irrelevant difficulty is present when aspects of the task that are extraneous to the focal construct make the test more difficult for some examinees. If it is found that participants answering the computerized test items engage in cognitive processes that are irrelevant to the construct of reading comprehension, such as processes that reflect working memory or spatial ability components, the construct validity of the computerized test may come into question. The present study examined the issue of the comparability of computerized and paper-and-pencil tests from a cognitive perspective, using verbal protocols, rather than psychometric methods, as its primary mode of inquiry. Reading comprehension items were the focus of this study, because these items are frequently used on achievement and aptitude tests, and because previous research has suggested that these items are more susceptible to mode effects than other item types (Mazzeo & Harvey, 1988).

Theoretical Foundation

Reading comprehension test items on the Graduate Record Examination (GRE) were designed to assess an examinee's ability to read with understanding, insight, and discrimination (Educational Testing Service [ETS], 1996). When these items are presented on the computer, there is the possibility that working memory and spatial ability are also assessed. These additional constructs, which use cognitive resources that would ordinarily be allocated to comprehending the passage, may be introduced due to the difficulty reading text from a computer screen, the inability to underline or mark text in the passage, and the inability to see the entire passage and all of the test questions at one time.

The Difficulty Reading Text from a Computer Screen

According to information-processing models of reading, the reading process begins when the eye fixates upon the words on the page or the computer screen (Samuels & Kamil, 1984). There is some evidence that reading text from a computer screen has negative effects on comprehension, due to visual fatigue and distractions introduced by the new mode of delivery (Daniel, 1983; Heppner, Anderson, Farstrup, & Weiderman, 1985). According to LaBerge and Samuels' (Samuels & Kamil, 1984) information processing model of reading, there are two major tasks that are performed when we read: decoding and comprehension. Both of these tasks require attention. For most skilled readers, the decoding process is automatic, leaving all cognitive resources available for comprehension. However, there are some circumstances which require additional amounts of attention, such as when unfamiliar words are encountered or when words are printed in an unfamiliar typeface (Samuels & Kamil, 1984). The difficulties reading text from the computer screen may require more attentional resources allocated to decoding, which take away resources allocated to comprehending the text.

The Inability to Underline or Mark Text in the Passage

An important activity in reading comprehension is finding the main ideas in the text and making certain that these ideas are remembered, or can be found again later if needed (Pressley & Afflerbach, 1995). Once main ideas are identified, readers often find it important to flag them, either verbally or with the use of physical aids (i.e., underlining or highlighting). In their review of the research on study strategies, Anderson and Armbruster (1984) reported that although

underlining has been found to be no more effective than other studying techniques, several studies have found that students who underline have a greater recall of the information that was studied. Anderson and Armbruster believe that the primary facilitative effect of underlining occurs due to the amount of processing required to make the decision about what to underline. In accord with this hypothesis, examinees who underline text while reading the passage on a paper-and-pencil reading comprehension test may process the text more thoroughly than examinees taking a computerized test who are unable to underline information. In addition, examinees taking a computerized test may have a greater working memory load to remember the important information in the passage because they cannot underline or mark this information.

The Inability to See the Entire Passage at One Time

Reading comprehension tests presented on the computer often include long passages that are not entirely visible on one screen. To read the entire passage, examinees must scroll or page through the text. There is evidence to suggest that readers establish a visual memory for the location of items within a printed text based on their spatial location both on the page and within the document (Rothkopf, 1971, as cited in Dillon, 1992). This memory is supported by the fixed relationship between an item and its position on a given page. Scrolling may weaken a reader's visual memory, which may affect the reader's ability to search for and locate information in the passage (Dillon, 1992; Haas & Hayes, 1986).

Furthermore, in the process of scrolling through text, sentences are often split across screens, requiring the reader to remember the information in the first part of the sentence while paging or scrolling to reveal the rest of the sentence (Dillon, 1992). This situation may cause what has been termed the Split Attention Effect (Chandler & Sweller, 1991; Sweller & Chandler, 1994; 1991). According to the Split Attention Effect, the requirement to mentally integrate noncontiguous material imposes an unnecessary and heavy load on working memory. Sweller and Chandler identified this effect while studying the effect of instructional diagrams when the text explaining the diagrams was located on a separate page, and comparing that to when the associated text was located on the same page as the diagram. They found that when both the diagram and the text were needed to understand the concept, the separation of the material imposed an extraneous working memory load which impaired learning.

The Split Attention Effect may also be present when examinees are required to read and integrate text located on separate computer screens. The cognitive resources required to

integrate material on separate screens may compete with the more meaningful processes of reading comprehension, such as constructing the main idea, making inferences, and identifying important information. If a reader's cognitive resources are already taxed, then he or she may have difficulty performing the tasks necessary to comprehend the text, resulting in a poorer understanding and memory of the text (Afflerbach, 1990).

The Inability to See All of the Questions at One Time

An added constraint introduced by a computerized test is the inability to see all of the questions at one time. This makes it more difficult for examinees to preview the questions before reading the passage, choose which items to answer first and which items to return to later, and check the pattern of their responses. Examinees taking reading comprehension tests are often focused on the goal of finding the correct answers to the test questions rather than actually understanding the passage (Farr, Pritchard, & Smitten, 1990; Sternberg, 1991). With this goal, many examinees choose to read the questions before reading the passage, using the questions to guide their reading. In fact, several test coaching companies advocate this strategy to their clients.

A "questions-first" strategy may have several cognitive benefits. It may reduce working memory load by limiting what the reader must attend to; it could aid examinees in their construction and integration of propositions by alerting them, a priori, to what information is important; and it could activate relevant schemas and scripts in the examinee's long-term memory (Bishop & Frisbie, 1998). Another frequently-used test-taking strategy is skipping and returning to questions. Most examinees desire the option to choose which questions to answer first and which to return to later, and it has been consistently found that a substantial proportion of examinees change answers to at least some questions (Vispoel, Hendrickson, Bleiler, Widiatmo, Sharairi, & Ihrig, 1999).

In the present study, it was hypothesized that the four constraints introduced by computerized reading comprehension tests - the difficulty reading text from the screen, the inability to underline or mark text, the inability to see the entire passage at one time, and the inability to see all of the questions at one time - introduced a working memory load for participants taking the computerized test. This was expected to result in different processes or a

different frequency of processes on the two test modes. The hypotheses of this study are summarized in Table 1.

Table 1
Hypotheses of the Study

Cognitive Processes During Initial Reading of the Passages

1. Reading Comprehension Processes. Participants taking the paper-and-pencil test will have a significantly greater frequency of utterances indicating comprehension of the passage, including rereading, paraphrasing, identifying important information, making inferences, and integrating text, than participants taking the computerized test.
2. Monitoring Location. Participants taking the computerized test will have a significantly greater frequency of utterances indicating monitoring of their location in the passage, than participants taking the paper-and-pencil test.

Cognitive Processes While Answering the Test Questions

3. Monitoring Processes. Participants taking the computerized test will have a significantly greater frequency of utterances indicating a lack of understanding, and a lack of recall for both content and location of information in the passage than participants taking the paper-and-pencil test.
4. Re-evaluation of Answer Choices. Participants taking the computerized test will re-evaluate the same answer choices significantly more frequently than participants taking the paper-and-pencil test.
5. Reading All Five Answer Choices Before Evaluating Them. Participants taking the computerized test will read all of the answer choices before evaluating any of them significantly more frequently than participants taking the paper-and-pencil test.

Search Strategy

6. Frequency of Searches. Participants taking the computerized test will engage in a significantly greater number of searches to find information in the passage to answer the test questions than participants taking the paper-and-pencil test.
7. Duration of Searches. The searches of participants taking the paper-and-pencil test will be significantly shorter in duration than the searches of participants taking the computerized test.
8. Characteristics of Searches. The searches of participants taking the paper-and-pencil test will result significantly more frequently in locating the information relevant to answering the test questions than the searches of participants taking the computerized test.
9. Use of Prior Work During Initial Reading. Participants taking the paper-and-pencil test will refer significantly more frequently to information that they underlined or marked during initial reading of the passages than participants taking the computerized test.

Overall Test-Taking Strategy

10. Reading Questions First. Participants taking the paper-and-pencil test will read the questions before reading the passages significantly more frequently than participants taking the computerized test.
11. Skipping and Returning to Questions. Participants taking the paper-and-pencil test will return to questions significantly more frequently than participants taking the computerized test.

Methodology

Participants

Forty-eight juniors and seniors (35 females and 13 males) from a large northeastern public university participated in this study. Participants were randomly divided into four groups, with twelve participants per group. Groups A and C took the computerized test (CT) first, followed

BEST COPY AVAILABLE

by the paper-and-pencil (PP) test. Groups B and D took the pencil-and-paper test first, followed by the computerized test. Groups A and B were asked to think aloud as they answered the reading comprehension items, while Groups C and D completed the items silently. All participants were administered the two passages in the same order; only the mode of administration differed. The four groups did not differ significantly in their mean reported SAT-Verbal score, F (3, 40) = 1.191, p = .325, in their experience taking any type of computerized test, χ^2 (3, N = 48) = 1.07, p = .785, or in their experience taking the GRE, χ^2 (3, N = 48) = .273, p = .965.

Materials

The reading comprehension items used in this study were taken from the Educational Testing Service (ETS) Graduate Record Exam (GRE) - General Test Big Book (1996), which presents retired GRE items used on actual tests administered between 1984 and 1994. Table 2 presents the characteristics of the GRE reading passages and test items that were used in this study. Two long passages (those which require scrolling in the computerized test), each consisting of 55 lines and seven corresponding test items, were selected from the Big Book. As this book provides information on item difficulty (p-values, or the percentage of examinees answering each question correctly), passages were chosen with relatively large mean p-values (i.e., the easiest passages), so that the task difficulty would not interfere with the process of thinking aloud, as has been suggested in some research (e.g., Afflerbach & Johnston, 1984).

Table 2
Characteristics of GRE Reading Passages and Test Items Used in the Study

	Passage 1 (Ragtime)	Passage 2 (Griffith)
Number of Lines	55	55
Number of Test Items	7	7
P-Values (Min., Max., Mean)	.56 to .83 (.70)	.40 to .92 (.69)
Sum of P-Values	4.93	4.83

Computerized Test. The computerized test used in this study was delivered on a web page that resembled the screen presenting reading comprehension items from ETS's GRE

PowerPrep (1997) software for the general test. An introductory screen with the instructions for completing the test was presented first. This introductory screen included instructions to click an icon to proceed to the passage and questions once the instructions had been read. The main test screen displayed the reading passage in a scrollable text field on the left half of the screen, with one item at a time presented on the right half of the screen. The text field containing the reading passage displayed 25 lines of text at one time. Participants indicated their answer to an item by clicking the mouse to darken a circle corresponding to their answer choice. Only one answer choice could be marked at a time. Unlike the actual PowerPrep software, the computerized test used in this study did not include icons at the bottom of the screen to enable examinees to review and mark items, access the time elapsed, access a help menu, or exit the test. Participants were provided with icons at the top left corner of the screen to enable them to move forward and backwards through the items corresponding to a passage. Participants were instructed that they could skip and return to items, and change their answers if they wished to do so.

Procedures

A brief interview was conducted with each subject at the beginning of the session, to ascertain their level of experience, familiarity, and comfort with computers and computerized tests, their experience taking the GRE or practice items, and their previous verbal SAT or GRE scores. Before taking the tests, all participants were asked whether they were comfortable using a computer and mouse to scroll through text and click on items. Because all participants indicated computer familiarity, it was not necessary to give participants a tutorial on the use of the computer.

Following the pre-experiment interview, participants in the two experimental groups were asked to think aloud as they answered the reading comprehension items on the computer and with paper-and-pencil. The instructions and warm-up tasks given to participants were adapted from the text suggested by Ericsson and Simon (1993). At the end of the session, participants in the experimental and control groups were given a brief post-experiment interview to learn about their perceptions and strategies taking the computerized and paper-and-pencil items, and whether or not they thought that their performance was better on either of the two test modes. Because the ETS Big Book is available to the public, it was possible that participants had prior exposure to the

experimental passages and test items. Therefore, during the post-experiment interview, participants were asked whether they had ever taken a practice GRE test, whether they used the Big Book, and whether they recalled any of the items they took during the experiment. If participants indicated that they used the Big Book, they were asked to peruse the practice tests containing the items they took during the experiment, and to indicate whether other items on those practice tests seemed familiar to them. Only one subject indicated ever using the Big Book, and this subject said that she did not remember either of the reading passages that were included in the experiment. Therefore, data from all 48 participants were included in the study.

Coding of The Verbal Protocol Data

The verbal protocols were transcribed verbatim, then divided into utterances corresponding to pauses in the protocols. As participants read the passages verbatim, each sentence was coded as a separate utterance. Three levels of coding were used to infer participants' cognitive processes and strategies as they read the passages and answered the test questions: the utterance level, the passage/question level, and the test level.

Coding at the Utterance Level. Coding of the utterances was guided by several well established accounts of the cognitive processes that occur during reading, including Pearson, Roehler, Dole, and Duffy (1992) and Pressley and Afflerbach (1995). The utterances during initial reading of the passages were coded into the following categories: reading (and rereading), paraphrasing, identifying important information, monitoring understanding, recalling (content and location of information), making inferences, integrating text, executive processes (i.e., stating strategies, monitoring progress), and using physical aids (i.e., underlining or writing).

Because it has been documented that individuals taking reading comprehension tests engage in unique behaviors that differ from other reading tasks (Cohen, 1986; Farr, Pritchard, and Smitten, 1990), the cognitive processes specified by reading researchers did not account for all of the behaviors exhibited by the participants in this study. Therefore, several additional codes were used, including reading and rereading question stems and answer choices, evaluating and re-evaluating answer choices, and selecting answer choices. Behaviors related to searching the passage were also coded, based on the research on document search (Guthrie, 1988). These behaviors included initiating searches, monitoring searches, evaluating searches, evaluating the

relation between text and answer choices (matching), and using prior work. A navigation code was used to indicate when participants moved from one question to another. The Appendix includes a full description and example of each coding category.

Coding at the Passage and Question Level. To code at the passage level, all utterances made by participants while they initially read the passage were kept together, and the content and sequence of utterances was examined. Coding at the passage level was used to infer whether participants engaged in selective reading or skimming of the passage, and whether participants skipped to the questions in the midst of reading the passage. Similarly, all utterances made by participants while they answered each of the test questions were kept together. Coding at the question level was used to infer participants' test-taking strategies, which included searching the passage for information to answer the questions, and reading all of the answer choices before evaluating them.

The coding of searches at the question level was guided by Guthrie's (1988) cognitive model for document search. Following Guthrie's model, each time participants searched the passage, the goal of the search and the category of the search was coded. In instances when participants overtly stated their intention to search the passage for information to answer a question, an "initiate search" code was present at the utterance level. The goal of the search (e.g., a word, phrase, or idea) was coded when participants overtly stated the information they were searching for. However, in most cases, participants did not overtly state their intention to search the passage, nor did they state the goal of their search. In these cases, searches were identified by utterances coded as reading, skimming, or paraphrasing sentences from the passage.

The category of a search was determined by examining what sentences or paragraphs were read or skimmed during the search. The sentences or paragraphs that contained the information necessary to answer each test question was identified by the researcher and verified by a second individual. Items asking participants to identify the main idea of the passage were excluded from this analysis, because the information to answer these questions could not be confined to a specific portion of the passage. The number of utterances between the initiation of a search and end of the search was recorded as a measure of the duration of the search.

Coding at the Test Level. The final level of coding was at the test level. At this level, participants' entire protocol generated during each test was coded for one of the following overall strategies: 1) reading the entire passage before reading any of the questions; 2) reading some of the passage, but skipping to one or more questions before finishing the passage; 3) reading all seven questions before reading the passage; 4) reading and answering one question at a time, and searching the passage for information to answer each question; 5) reading one or more question stems, then reading the passage, then returning to the questions; or 6) another strategy. It was expected that participants taking the paper-and-pencil test would read and/or answer some or all of the questions before reading the passage more frequently than those taking the computerized test.

Inter-Coder Reliability. Inter-coder reliability was established for coding at the utterance level, and was based on 1,977, or approximately 25 percent of the total number of utterances from six randomly selected participants from each group. A graduate student in psychology recoded the utterances, and inter-coder reliability was established using Cohen's Kappa (Bordens & Abbot, 1991). Inter-coder agreement was 86 percent, and Cohen's Kappa was .80. The utterances which were coded differently were discussed among the two coders until full agreement was reached.

Results

Table 3 displays the means and standard deviations of the test scores and time spent on both test modes for each group. A three-way analysis of variance was conducted on test scores with group (experimental and control) and order of administration (computerized-first or paper-and-pencil-first) as between-subject factors and test mode as a repeated factor. There were no significant main effects, nor were any interactions significant. A comparison of performance within groups revealed that the two experimental groups did slightly better on the second passage, while the two control groups did slightly better on the first passage, regardless of mode.

Table 3
Means and Standard Deviations of Scores (Number of Questions Correct) and Minutes Spent on Computerized and Paper-and-Pencil Tests

	Group	N	Computerized Test		Paper-and-Pencil Test	
			Score	Time	Score	Time
A.	Experimental (CT-first)	12	5.00 (1.65)	16.00 (4.02)	5.42 (1.24)	13.55 (4.46)
B.	Experimental (PP-first)	12	5.33 (1.23)	12.73 (2.83)	5.08 (1.68)	13.36 (3.85)
C.	Control (CT-first)	12	5.17 (1.34)	11.00 (3.95)	4.92 (1.16)	8.50 (2.32)
D.	Control (PP-first)	12	4.92 (1.00)	9.67 (1.78)	5.08 (1.16)	9.27 (3.20)

A three-way analysis of variance was also conducted on time spent on the two tests, with group and order of administration as between-subject factors and test mode as a repeated factor. There was a significant main effect for group (F (1,41) = 22.78, $p < .05$), but not for the order of administration. The repeated factor was statistically significant (F (1,41) = 6.15, $p < .05$), as was the interaction between this factor and the order of administration (F (1,41) = 8.26, $p < .05$). These results show that the experimental groups spent a significantly greater amount of time than the control groups on both the computerized and paper-and-pencil tests. These findings are consistent with the literature stating that thinking aloud slows down the reading process and increases the time needed to complete a task (Ericsson & Simon, 1993; Pressley & Afflerbach, 1995; Bereiter & Bird, 1984).

The significant interaction between test mode and order of administration reflects differences in the time spent on the first and second tests, or a practice effect. Groups A and C, who took the computerized test first, spent more time on the computerized test than on the paper-and-pencil test. Group D, who took the paper-and-pencil test first, spent an approximately equal amount of time on both tests. However, Group B spent more time on the paper-and-pencil test. It appears that most participants took longer to complete the first test than the second test,

regardless of mode. The interaction of time and group, and the three-way interaction of time, order, and group were not statistically significant.

The next sections present the results addressing the hypotheses that a working memory load on the computerized test affected the cognitive processes and strategies used by participants while reading the passage and answering the test questions. Although the original intention was to combine the data for the two tests, there was evidence that the two tests differed in perceived difficulty and interest-level.[1] As shown earlier, participants did better on the second test and spent more time on the first test, regardless of test mode. In addition, many participants reported after the experiment that the first test was more difficult and less interesting than the second test. In light of these findings, there was concern that if the data for the two tests were combined, mode effects would be confounded with the difficulty and interest level of the passages.

Because the evidence of differential difficulty and interest-level for the two tests was only anecdotal, a small follow-up study was conducted. Five graduate students took the two tests with paper-and-pencil in a counterbalanced order (three took the Ragtime test first and two took the Griffith test first), and were then asked to indicate which set of questions was more difficult, which passage was easier to understand, and which passage they found more interesting. Four of the five students indicated that the questions associated with the Ragtime passage were more difficult, and all five students indicated that the Griffith passage was both easier to understand and more interesting. Due to these findings, the decision was made to analyze the two tests separately.

Cognitive Processes During Initial Reading of the Passages

The analyses conducted to compare the cognitive processes during initial reading of the passages were based on coding at the utterance level. The first hypothesis of this study was that there would be a significantly greater frequency of utterances reflecting comprehension of the text during initial reading of the passages on the paper-and-pencil test than on the computerized test. For the purposes of this study, reading comprehension processes included any activity other than

1 Although the first test was perceived as more difficult than the second test, in the GRE test-taking population, the questions associated with the two passages were of similar difficulty. The estimated mean scores of the National GRE sample (ETS, 1996) are 4.93 for the first test and 4.83 for the second test.

verbatim reading of the passage that indicated that the participant was attempting to comprehend the text. These processes included rereading, paraphrasing, making inferences, monitoring understanding, integrating text,[2] and identifying important information (either verbally or by using physical aids). The use of physical aids was coded when participants overtly stated their intention to use physical aids, or when underlining or other writing was found on the passage.

Executive processes included utterances that reflected monitoring location, monitoring progress, and stating strategies.[3] The other processes that were coded included verbatim reading or skimming from the passage, directions, or questions, and other miscellaneous processes. Table 4 shows the distribution of processes for each test mode on the first and second tests. Chi-square tests of association were used to test the null hypothesis that the cognitive processes in each category (reading comprehension processes and executive processes) were independent of test mode. An alpha level of .01 was used for each of these tests. Table 5 shows each test's statistic and probability level.

Reading Comprehension Processes. On the first test taken by both experimental groups (Ragtime), participants taking the paper-and-pencil test had a significantly greater frequency of utterances reflecting reading comprehension processes than participants taking the computerized test. Identifying important information was the most frequently used reading comprehension process on the paper-and-pencil test, and all but one of the utterances indicating the identification of important information reflected the use of physical aids. Because using physical aids was not as convenient for participants taking the computerized test (these participants were not able to write on the passage, but they were given scratch paper to use as they wished), an analysis was also conducted removing this process.

When identification of important information was removed from the analysis, the difference in initial reading processes on the computerized and paper-and-pencil tests was no longer statistically significant, although reading comprehension processes still accounted for a

2 Because there was only one utterance for integrating text, this process was excluded from the analyses.

3 Although executive processes such as monitoring progress and verbatim reading and skimming are often considered part of the reading comprehension process, in this study reading comprehension processes were considered those processes in which the participant was actively engaged in determining the meaning of the text.

greater percentage of the total utterances during initial reading of the passage on the paper-and-pencil test than on the computerized test. There was a trend for participants taking the paper-and-pencil test to reread the text slightly more frequently than those taking the computerized test. Similarly for the second test, when identifying important information was included in the analysis, participants taking the paper-and-pencil test engaged in reading comprehension processes significantly more frequently than participants taking the computerized test. But, when identifying important information was removed, the difference was no longer statistically significant. In both test modes, participants made very few utterances reflecting inference-making, paraphrasing, integrating text, and monitoring understanding. The vast majority of utterances reflected verbatim reading of the passage, which suggests that participants either did very little other than a surface-level reading of the text during initial reading of the passage, or did not report many of their cognitive processes as they were reading.

Executive Processes. The second hypothesis with regard to initial reading of the passages was that participants taking the computerized tests would have a significantly greater frequency of utterances indicating monitoring of their location while they initially read the passage. The chi-square results indicated that the difference in the frequency of executive utterances was significant for both tests. On both the first and second tests, participants taking the computerized test monitored their location and monitored their progress slightly more frequently, but participants taking the paper-and-pencil test stated strategies much more frequently. Many of the strategy utterances made by participants on the paper-and-pencil tests reflected the intention to underline or mark text during initial reading. Because this strategy was not available to participants taking the computerized tests, this explains the fewer strategy utterances for this test mode.

Table 4
Cognitive Processes During Initial Reading of the Passage on the Computerized and Paper-and-Pencil Tests

Processes		Total Number of Utterances			
		First Test (Ragtime)		Second Test (Griffith)	
		CT	PP	CT	PP
Reading Comprehension Processes	Rereading	7	26	34	32
	Paraphrasing	1	7	5	2
	Making Inferences	3	5	5	4
	Monitoring Understanding	4	6	0	1
	Integrating Text	0	0	1	0
	Identifying Important Information				
	Verbally	0	1	0	3
	Using Physical Aids	0	73	9	35
Executive Processes	Monitoring Location	5	1	5	0
	Monitoring Progress	5	1	5	2
	Stating Strategy	5	36	9	28
Verbatim Reading, Skimming, and Miscellaneous Processes	Verbatim Reading from Passage	208	220	210	221
	Reading Directions	12	12	12	12
	Skimming	0	7[a]	0	0
	Reading Questions/Answers	1	13	6	2
	Miscellaneous	9	22	30	13
Total Utterances During Initial Reading		256	353	327	308
Total Reading Comprehension Utterances		**15**	**118**	**54**	**77**
Total with Physical Aids Removed		15	45	45	42
Percentage Accounted for by Reading Comprehension Processes		**5.8%**	**33.4%**	**16.5%**	**25.0%**
Percentage with physical aids removed		5.8%	12.7%	13.8%	13.6%

Note. Some utterances were coded into more than one category; therefore, the frequencies reported in the table do not equal the total number of utterances.
[a]These seven utterances came from one participant.

Table 5
Pearson Chi Square Statistics for Cognitive Processes During Initial Reading

Process	First Test				Second Test			
	n	df	χ^2	p	n	df	χ^2	p
Reading Comprehension Processes (rereading, paraphrasing, identifying important information, making inferences, monitoring understanding, and integrating text)	133	4	26.41	< .01	130	4	16.48	< .01
Reading Comprehension Processes with Identifying Important Information Removed	59	3	2.75	.432	83	3	2.16	.539
Executive/Monitoring Processes (monitoring location, monitoring progress, and stating strategies)	53	2	23.15	< .01	49	2	14.29	< .01

Cognitive Processes While Answering the Test Questions

It was hypothesized that due to a working memory load on the computerized test, participants taking this test would: 1) have a significantly greater frequency of utterances indicating a lack of understanding and a lack of recall for both content and location of information in the passage; 2) re-evaluate answer choices more frequently; and 3) read all of the answer options before evaluating any of the options more frequently. The analyses to address the first two hypotheses were based on coding at the utterance level, while the analysis to address the third hypothesis was based on coding at the question level. Table 6 displays the frequency of cognitive processes while participants answered the questions on the computerized and paper-and-pencil tests.

To test the first two hypotheses with regard to answering the test questions, chi-square tests of association were conducted on the frequency of utterances in five categories: evaluating answer choices, executive processes, search processes, reading comprehension processes, and using physical aids. The coding of utterances within each of these categories was mutually exclusive, therefore meeting the chi-square test's assumption of independence. A Bonferroni-adjusted alpha level of .01 was used for each of the five tests. Table 7 displays each test's statistic and probability level.

BEST COPY AVAILABLE

Table 6
Cognitive Processes While Answering the Questions on the Computerized and Paper-and-Pencil Tests

Processes	Total Number of Utterances			
	First Test (Ragtime)		Second Test (Griffith)	
	CT	PP	CT	PP
Evaluating Answer Choices				
Initial Evaluation	206	178	217	253
Re-evaluation	63	14	67	77
Evaluating Selected Choices	7	3	8	5
Crossing Out Incorrect Choices	6	134	7	173
Marking Choices	0	15	2	24
Executive/Monitoring Processes				
Recalling Content				
Negative	5	4	2	10
Positive	29	19	20	37
Recalling Location				
Negative	4	7	3	3
Positive	3	4	2	2
Stating Strategy	46	37	23	21
Monitoring Progress	30	19	28	35
Search Processes				
Initiating Searches	41	28	25	33
Monitoring Searches	16	10	8	14
Reading/Skimming/Paraphrasing Passage	327	257	286	232
Evaluating Searches	22	9	3	17
Matching Text with Answer Choices	14	9	5	22
Reading Comprehension Processes				
Rereading Questions and Answer Choices	159	145	137	111
Making Inferences	33	16	26	46
Integrating Text	10	6	1	0
Identifying Important Information				
Verbally	10	1	3	3
Underlining/Marking	0	24	0	26
Writing	4	4	7	8
Monitoring Understanding				
Negative	15	5	3	4
Positive	0	0	3	1
Reading Questions and Answer Choices	504	476	355	395
Selecting Answer Choices	90	88	94	89
Miscellaneous/Other Processes	260	188	250	254
Total Number of Utterances While Answering Questions	1,864	1,493	1,514	1,646

Note. Some utterances were coded more than once; therefore, the frequencies reported in the table do not equal the total number of utterances.

Table 7
Chi-Square Statistics for Cognitive Processes During Question Answering

Process	First Test				Second Test			
	n	df	χ^2	p	n	df	χ^2	p
Evaluating Answer Choices (initial evaluation and re-evaluation)	461	1	20.95	< .01	614	1	.01	.94
Executive Processes (recalling content, recalling location, stating strategies, and monitoring progress)	207	3	2.99	.39	186	3	5.22	.16
Search Processes (initiating search, monitoring search, reading/skimming/paraphrasing passage, evaluating search, and matching)	733	4	3.21	.52	645	4	28.75	< .01
Reading Comprehension Processes (rereading, making inferences, integrating text, identifying important information, and monitoring understanding)	432	4	15.77	< .01	379	4	23.99	< .01
Reading Comprehension Processes with Identifying Important Information Removed	389	3	7.44	.06	332	3	9.18	.03
Physical Aids (marking text, writing, crossing out or marking answers)	187	3	34.03	< .01	247	3	43.80	< .01

Evaluating Answer Choices. On the first test, participants taking the computerized test re-evaluated answer choices significantly more frequently than participants taking the paper-and-pencil test. However, there was no difference in the frequency of re-evaluating answer choices on the second test.

Executive Processes. On both the first and second tests, there was no difference in the frequency of executive processes on the two test modes.

Search Processes. There was no difference in the frequency of search processes on the first test. On the second test, however, there was a significant difference in the frequency of search processes on the two test modes. Participants taking the paper-and-pencil test initiated, monitored, and evaluated their searches, and matched text in the passage to the answer choices more frequently, while participants taking the computerized test read, skimmed, or paraphrased the passage more frequently.

Reading Comprehension Processes. On both the first and second tests, there was a significant difference in the frequency of reading comprehension processes on the two test modes. On the first test, participants taking the computerized test monitored their lack of understanding of the passage slightly more frequently than participants taking the paper-and-pencil test. Participants taking the computerized test also made inferences more frequently, while those taking the paper-and-pencil test identified important information more frequently. On the second test, there was no difference in the frequency of comprehension monitoring, but participants taking the computerized test reread question stems and answer choices more frequently, while those taking the paper-and-pencil test made inferences and identified important information more frequently.

Use of Physical Aids to Identify Important Information. As expected, on both the first and second tests, there was a significantly greater use of physical aids on the paper-and-pencil test.

Reading All Five Answer Choices Before Evaluating Them. The third hypothesis with regard to answering the test questions was that participants taking the computerized tests would read all of the answer choices before considering them more frequently than participants taking the paper-and-pencil tests. Coding at the question level indicated whether a participant read all five answer choices before evaluating them for each of the seven questions per test. The means were calculated by summing the number of times participants read all five answer choices for each test mode and dividing by 84, which is the number of participants (12) times the number of questions (7). An independent t test was used to test the difference in the means.

On both the first and second tests, participants taking the computerized test did read all five answer choices more frequently than participants taking the paper-and-pencil test. The mean difference for the first test was marginally significant at a Bonferroni-adjusted alpha level of .025 (t (165.05) = 2.07, p = .04), and the mean difference for the second test was statistically significant (t (160.28) = 2.29, p = .024). The variances for both tests were significantly different as indicated by Levene's Test for Equality of Variances; therefore, the t tests were based on separate variance estimates.

Search Strategy

As expected, the participants in this study engaged frequently in a search strategy as they attempted to locate information to assist them in answering the test questions. With regard to participants' search strategy, it was hypothesized that those taking the computerized test would

initiate a greater number of searches, and that the searches would be longer in duration, less targeted, and less frequently related to prior work during initial reading of the passages. The analyses to address these hypotheses were based on coding at both the utterance level and the question level.

As described earlier, coding at the utterance level revealed no significant difference in the frequency of search processes on the first test taken by participants. On the second test, there was a significant difference in the frequency of search processes for the two test modes. A comparison of the chi-square expected values and residuals for the search processes revealed that participants taking the computerized test read, skimmed, and paraphrased the passage more frequently than those taking the paper-and-pencil test, while participants taking the paper-and-pencil test initiated, monitored, and evaluated their searches, and matched text from the passage and answer choices more frequently than those taking the computerized test.

Further analyses of participants' search strategies were conducted, based on coding at the question level. At this level, the number of searches initiated by each participant for a given question, and the duration of each search was coded. The search duration is the number of utterances between the initiation of a search and the end of a search, and is considered more informative than the number of searches due to individual differences in the frequency of searches. For example, one participant may have had several short searches, while another may have had fewer longer searches. Nevertheless, the search duration for the two participants may have been the same. The search duration is a measure of how much time (number of utterances) participants spent searching the passage for information to answer the questions, regardless of the number of searches that were coded.

Coding at the question level also captured the characteristics of each search, including whether the participant: located information relevant to selecting the correct answer choice, located information relevant to rejecting the incorrect answer choices, located information that was irrelevant to the test questions, located information that could potentially lead them to select an incorrect answer (i.e., located misleading information), and searched the same portion of text more than once for a given question (i.e., repeated a search).

Number and Duration of Searches. Independent t tests were used to test the difference in the means, using an alpha level of .01. None of the differences were statistically significant. Although there were no differences in the number and duration of searches on the two test modes overall, it was speculated that mode effects might be apparent for certain types of test items and not for others. For example, examinees taking a computerized test might search more frequently to find the answer to main idea questions, due to a poorer overall understanding of the passage. On the other hand, examinees taking a paper-and-pencil test might search more frequently to find the answer to supporting idea questions, because the information to answer this type of question is explicitly stated in the passage, and there is evidence that it is easier to locate specific information on paper than on a computer screen.

To test this hypothesis, the questions used in this study were categorized by type, as practiced by item writers at Educational Testing Service (K. Cureton, personal communication, December 17, 1999). Educational Testing Service uses six item categories for its reading comprehension test items: main idea/main purpose, supporting idea, inference, application, evaluation, and style. For the purposes of this study, it was of primary interest to distinguish between the type of question for which the information to answer the question was explicitly mentioned in the passage and those for which the information had to be deduced or inferred. Therefore, the ETS item types were collapsed into four broader categories: main idea, supporting idea, inference and other (application, evaluation, and style).

Main idea questions require a global understanding of the major purpose or focus of the passage. Supporting idea questions test the ability to identify and understand ideas explicitly mentioned in the passage. These questions potentially could be answered without an overall or complete comprehension of the passage. Inference questions test the ability to draw inferences about ideas or statements in the passage, and to understand implications of these ideas that are not explicitly stated. The remaining questions require examinees to apply elements of the passage to situations or problems outside of the passage (application), to identify and evaluate the logical structure of the passage or the author's methods of argument or persuasion (evaluation), or to identify the tone and/or style of the passage (style). Similar to main idea and inference questions,

it was presumed that these questions require higher-order thinking skills, which would rely on at least a partial comprehension of the passage.

Table 8 shows the total number, mean number, and mean duration of searches on the computerized and paper-and-pencil tests, by question type. Independent *t* tests were used to compare the mean number and mean duration of searches on the two test modes. None of the differences were statistically significant at an alpha level of .01. Because there were only one to three items in each category, and the low power of the study may have obscured meaningful effects, a post-hoc analysis of the effect size of these differences was conducted.[4] This analysis indicated a small effect for the difference in search duration for supporting idea questions on the first test ($\omega^2 = .01$), and a medium effect for the difference in search duration for application, evaluation, and style questions on both the first and second tests ($\omega^2 = .12$ and $\omega^2 = .08$, respectively). The effect sizes for the remainder of the differences were less than .01. This suggests that mode differences may exist for the application, evaluation, and style questions, but that the low power of this study precluded finding statistically significant differences.

Search Characteristics. Because search is a very prevalent test-taking strategy for paper-and-pencil reading comprehension tests (Farr, Pritchard, & Smitten, 1990), perhaps the mode difference is more apparent in the nature of searches than in the number or duration of searches. For example, examinees taking paper-and-pencil tests might be more efficient and better targeted in their search strategy. That is, they might be more likely to locate the information relevant to the test questions, and they might be less likely to search irrelevant sections of the passage and search the same portion of text more than once per question.

4 Omega squared (ω^2) was used as an estimate of effect size. The effect sizes were categorized using the criteria described by Cohen (as cited in Keppel, 1991, p. 66). An ω^2 of .01 indicates a "small effect," an ω^2 of .06 indicates a "medium effect," and an ω^2 of .15 indicates a "large effect."

Table 8
Number and Duration of Searches on the Computerized and Paper-and-Pencil Tests, by Question Type

Test	Question Type	Total Number of Searches		Mean Number of Searches		Mean Search Duration	
		CT	PP	CT	PP	CT	PP
First	Main Idea (Question 1)	9	4	0.8	0.4	6.5	4.7
	Supporting Idea (Questions 2 & 3)	43	49	2.0	2.2	6.3	9.1
	Inference (Questions 5 & 6)	28	18	1.3	0.8	6.3	5.9
	Application/Evaluation (Questions 4 & 7)	35	21	1.6	0.9	8.5	4.8
Second	Main Idea (Question 1)	3	9	0.3	0.8	2.7	3.7
	Supporting Idea (Question 4)	22	27	2.0	2.5	8.4	6.6
	Inference (Questions 2, 3, & 5)	49	45	1.5	1.4	7.1	6.4
	Application/Style (Questions 6 & 7)	24	22	1.1	1.0	5.3	3.1

Table 9 shows the characteristics of participants' searches, by question type. The table shows the number of times participants located the information relevant to selecting the correct answer choice, the number of times participants located information relevant to correctly rejecting answer choices, the number of times participants located information that was irrelevant to the test questions, the number of times participants located information that could potentially lead them to select an incorrect answer (i.e., locating misleading information), and the number of times participants searched the same portion of text more than once for a given question (i.e., repeating a search).

Table 9
Characteristics of Searches on the Computerized and Paper-and-Pencil Tests, by Question Type

Test	Question Type[a]	Locates Relevant Information to Select Correct Choice		Locates Relevant Information to Reject Incorrect Choices[b]		Locates Irrelevant Information		Locates Misleading Information		Repeats Search	
		CT	PP	CT	PP	CT	PP	CT	PP	CT	PP
	Supporting Idea (Questions 2 & 3)	12	16	17	17	4	7	--	--	7	11
	Inference (Questions 5 & 6)	3	2	3	1	5	3	--	--	3	1
	Application/ Evaluation (Questions 4 & 7)	14	8	8	8	7	3	2	2	6	5
	TOTAL	29	26	28	26	16	13	2	2	16	17
	Supporting Idea (Question 4)	--	--	19	18	3	2	6	4	2	4
	Inference (Questions 2, 3 & 5)	9	10	19	16	6	4	8	6	4	4
	Application/ Style (Questions 6 & 7)	9	9	--	--	4	1	--	--	6	2
	TOTAL	18	19	38	34	13	7	14	10	12	10

Note. Dashes indicate that the information was not available for the items.
[a] Main idea questions were excluded from this analysis because the entire passage was considered relevant in answering these questions. [b] Participants could locate relevant information to reject choices a maximum of four times per question.

Independent *t* tests were used to compare the mean frequencies for each search characteristic, using an alpha level of .01. The results of this analysis revealed no significant differences in the characteristics of the searches on the computerized and paper-and-pencil tests when all seven questions were combined. When this analysis was conducted by question type, some differences in search characteristics on the computerized and paper-and-pencil tests were revealed, but none were statistically significant and the magnitude of the differences was very small.

Use of Prior Work During Initial Reading. A final hypothesis with regard to search strategy was that participants would refer to and use information they identified as important during initial reading of the passage more frequently on the paper-and-pencil test than on the computerized test. There were only five instances reflecting participants' use of prior work while answering the test questions. As expected, all five instances occurred on the paper-and-pencil tests (two on the first test, and three on the second test).

Overall Test-Taking Strategy

The final set of hypotheses of this study pertained to examinees' overall test-taking strategies on the computerized and paper-and-pencil tests. It was anticipated that participants taking the paper-and-pencil tests would read the questions before reading the passage more frequently than participants taking the computerized tests. There was no significant difference in the overall test-taking strategy of participants taking the computerized and paper-and-pencil tests on either the first or second test. Most participants used the same overall test-taking strategy regardless of test mode, that is reading the passage before answering the test questions.

It was also hypothesized that participants taking the paper-and-pencil tests would return to questions to review and/or change their answers more frequently than participants taking the computerized tests. A navigation code was used to indicate when a participant moved from one question to another. Participants who answered all seven questions and did not return to any of the questions had six navigation codes. The mean number of navigation codes was compared using an independent t test with an alpha level of .025.

On the first test, participants taking the computerized test had a mean of 6.6 navigation codes, compared to a mean of 6.5 for participants taking the paper-and-pencil test. On the second test, participants taking the computerized test had a mean of 7.0 navigation codes, compared to a mean of 7.4 for participants taking the paper-and-pencil test. Neither of these differences was statistically significant. It is shown that participants returned to questions more frequently on the second test, regardless of mode. In summary, there was no evidence that participants taking the paper-and-pencil test took advantage of the availability and proximity of the questions and read the questions before reading the passage, or returned to questions more frequently.

BEST COPY AVAILABLE

Summary of Findings

Table 10 presents a summary of the findings of this study in reference to the hypotheses. The asterisks indicate statistically significant differences in accord with the study's hypotheses, the checks indicate non-significant differences that are in the direction of the study's hypotheses, and the dashes indicate no differences. With regard to the cognitive processes during initial reading of the passages, the only significant finding was in the frequency of identifying important information on the two test modes. Participants taking the paper-and-pencil test identified important information in the passage quite frequently, while those taking the computerized test rarely did so. Participants taking the computerized test monitored their location in the passage more frequently than those taking the paper-and-pencil test, but the frequency was not large enough to suggest that this irrelevant process was dominant in the test.

Table 10
Summary of Findings

Hypotheses	First Test	Second Test
Cognitive Processes During Initial Reading of the Passages		
1. Greater frequency of reading comprehension utterances on PP.	*	*
When Identifying Important Information is Removed	–	–
2. Greater frequency of monitoring location utterances on CT.	✔	✔
Cognitive Processes While Answering the Test Questions		
3. Greater frequency of utterances reflecting lack of understanding of the passage on CT.	✔	–
Greater frequency of utterances reflecting lack of recall for the content and location of information in the passage on CT.	–	–
4. Greater frequency of re-evaluating answer choices on CT.	*	–
5. Greater frequency of reading all five answer choices before evaluating them on CT.	✔	*
Search Strategy		
6. Greater number of searches on CT.	–	–
7. Longer searches on CT.	–	–
8. Less targeted searches on CT.	–	–
9. Searches related to prior work on PP.	✔	✔
Overall Test-Taking Strategy		
10. Greater frequency of questions-first strategy on PP.	–	–
11. Greater frequency of returning to questions on PP.	–	–

Note. The asterisks indicate statistically significant mode differences in accord with the study's hypotheses, the checks indicate non-significant mode differences in the direction of the study's hypotheses, and the dashes indicate no mode differences.

With regard to the cognitive processes while answering the test questions, there was only slight evidence of a working memory load on the computerized test, and this evidence appeared for the first test but not the second test. Participants taking the first computerized test monitored their lack of understanding of the passage, re-evaluated answer choices, and read all five answer choices before evaluating them more frequently than those taking the paper-and-pencil test; however, only one of these three findings was statistically significant at the predetermined alpha levels. Finally, there was no evidence of any differences in search strategies or in overall test-taking strategies on the computerized and paper-and-pencil tests.

These findings suggest that computerized and paper-and-pencil tests may be more cognitively similar than originally thought. There was very little evidence that the difficulty of reading text from the computer screen, the inability to write on the passages and test questions, and the inability to see the entire passage and all of the questions at one time introduced construct-irrelevant variance into the test that affected participants' engagement in the construct-relevant behaviors. In fact, some of the findings indicate that computerized tests may encourage more construct-relevant behaviors than paper-and-pencil tests. This will be discussed in greater detail in the following section.

Discussion

Differences in Cognitive Processes During Initial Reading of the Passage

In this study, it was expected that participants taking the computerized test would experience a working memory load during initial reading of the passage, and that this would be reflected in a lower frequency of reading comprehension utterances and a higher frequency of monitoring location utterances. Although there was a significantly greater frequency of reading comprehension utterances on the paper-and-pencil tests, these were largely accounted for by the use of physical aids to identify important information in the passage. During their post-experiment interviews, many participants commented that physical aids are a predominant strategy when taking reading comprehension tests, and that they felt disadvantaged on the computerized test by not being able to write on the passage and test questions.

Although identifying important information is an important component of reading comprehension, there was no indication that participants taking the paper-and-pencil test engaged

more frequently in other reading comprehension processes such as paraphrasing, making inferences, monitoring their understanding, and integrating text. Furthermore, there was no evidence that the identification of important information led to a deeper processing of the text, nor was there evidence that this process facilitated participants' searches while they answered the test questions.

It was surprising to find that such a frequently-used strategy did not seem to produce any cognitive benefits. This suggests that examinees taking a paper-and-pencil test may have a false perception that having the ability to write on the passage improves their comprehension of the passage. On the contrary, underlining words and sentences in the passage may give examinees a false sense of security, in that they use the underlining as a substitute for more meaningful reading comprehension processes. On the other hand, examinees taking a computerized test may be compelled to process the text more deeply because they cannot rely on physical aids. If this is the case, the computerized tests might actually have better construct validity than the paper-and-pencil tests.

In this study, most participants did not engage in (or did not verbalize) reading comprehension processes when they read the passage initially, regardless of test mode. This may either be an intentional strategy, or it may be the result of difficulty thinking aloud during continuous reading. During the post-experiment interview, one participant reported that although she read the entire passage first, she did not expend a lot of effort to comprehend the passage because she knew she would have to return to the passage to answer the questions anyway. This supports the claim that examinees taking reading comprehension tests are usually focused on answering the questions, and engage in different processes than if they had the goal to learn or understand the material. Although the overall test-taking strategy of most of the participants in this study involved reading the passage first, perhaps this initial reading was only at a surface level, with the goal not to comprehend the passage, but to get a sense of the topic and location of information, so that searching to find the information to answer the questions was more effective.

The small percentage of reading comprehension utterances during initial reading may have also been due to participants' difficulty thinking aloud. During the course of reading, a few participants overtly stated this difficulty, and others admitted during the post-experiment

interview that they did not reveal all of their cognitive processes during initial reading. Several researchers have pointed out the potential for difficulty in the use of the concurrent think-aloud method during continuous reading, especially when the text is difficult (Afflerbach & Johnston, 1984; Bereiter & Bird, 1985). Future studies might employ a different think-aloud method, such as the "marked method," where participants are asked to read silently and think-aloud only at predetermined points in the text, rather than continuously during reading. Although the marked method does not provide data that are as complete as the concurrent method, this method might have been more effective in this study, especially in light of the finding that the first passage was difficult for participants to comprehend.

Differences in Cognitive Processes While Answering the Questions

In this study, there was slight evidence of a working memory load while answering the questions on the computerized tests, which was stronger for the first test than for the second test. The analysis of search frequency and search characteristics such as the frequency of locating information relevant to answering the questions revealed very few mode differences, and the magnitude of the differences was very small. Further research is needed to understand the relationship between test mode and question type and the effect on the strategies used to take reading comprehension tests.

Differences in Overall Test-Taking Strategies

Although it was expected that participants taking the paper-and-pencil test would read the questions before reading the passage more frequently because all of the questions were visible and easily accessible, this was not found to be the case. Nearly all participants used the same overall test-taking strategy on the two test modes which entailed reading the passage in its entirety before turning to the test questions. Although most of the participants read the passage first, several reported during the post-experiment interview that when they take reading comprehension tests under normal conditions, they usually read the questions first. Some said that they attended test review courses where they were taught to use this strategy. When asked why they used a different overall strategy during the experiment, a few said that when they are faced with a time limit, they usually choose the questions-first strategy, but because there was no time limit imposed during the experiment, they chose to read the passage first.

Several participants mentioned during the post-experiment interview that they would have liked to have seen all of the questions on the computerized test, and that they probably would have read some or all of the questions had they all been visible. On the computerized test, participants had the opportunity to view all of the questions, but in order to do so, they would have had to click on an arrow at the top of the screen and move through each of the question screens. Some participants said that they did not want to make the effort to go through all of the question screens, while a few other participants who had recently taken the computer adaptive GRE, which did not permit returning to questions, said that they assumed that the computerized test they took during the experiment had the same constraints.

It is very possible that the lack of a time limit in this study affected the overall strategies used by participants. A time limit may have led to the test-taking strategies predicted in this study. That is, more participants taking the paper-and-pencil test may have read the questions first, while the majority of participants taking the computerized tests may have continued to read the passage first. It was also hypothesized that participants taking the paper-and-pencil test would return to questions to review and/or change their answers more frequently than participants taking the computerized tests, again due to the visibility of all questions at once. In this study, there was no difference in the frequency of returning to questions for the two test modes. It is uncertain whether imposing a time limit would have changed these results.

There is some concern that reading comprehension tests may measure different constructs depending upon whether examinees read the questions first or whether they read the passages first (Bishop & Frisbie, 1998). The directions on most reading comprehension tests instruct examinees to read the passages first, implying that examinees should attempt to comprehend the passages before turning to the questions. However, as demonstrated in this study, examinees taking paper-and-pencil reading comprehension tests are usually focused on the questions and attempt to comprehend the passage only as much as is necessary to answer the questions. This leads to a strategy of searching the passage for information to answer the questions, which is quite different from the construct these tests were designed to assess. Since computerized tests present only one question at a time and make it more difficult for examinees to preview the items, computerized tests may encourage examinees to engage in behaviors that more closely resemble those that the

tests were designed to assess, that is, reading the passages first. This is another mode effect that may actually increase the construct validity of computerized tests.

Implications of Findings and Suggestions for Future Research

The verbal protocol method is extremely valuable for uncovering the cognitive processes and strategies used to perform a task. This method provides rich, authentic data which cannot be obtained through other more standard measures, such as test scores and questionnaires. The research questions that were posed by this study could not have been answered without the use of verbal protocols. However, it was recognized before the study began that the rich data would be obtained at the cost of statistical power. Some of the findings of this study indicate that computerized and paper-and-pencil reading comprehension tests may evoke different processes. However, with only 12 participants per group, many of the differences were not statistically significant. Therefore, it is difficult to make inferences with regard to the reliability of these differences. In this sense, the results of this study may be considered exploratory.

An unexpected, yet important finding of this study was the interaction between test mode and passage difficulty and interest level. The first test, which was less interesting and more difficult for participants, exposed more of the mode effects that were predicted in this study than the second test. The second test included a passage about the history of the cinema, a topic which evoked interest and some prior knowledge from participants. In fact, one participant stated during the post-experiment interview that as she read the passage, she was able to visualize some of the cinematic effects described in the passage based on movies she had seen.

There is considerable evidence that comprehension is best when the text is meaningful and relevant to the reader (Johnston, 1984; Pressley & Afflerbach, 1995). The activation and use of prior knowledge to interpret and relate text is an integral part of reading comprehension. Without some basis of prior knowledge, comprehension is difficult if not impossible. There is also evidence that prior knowledge facilitates search processes, as it directs attention to appropriate sections of the text, it facilitates extraction of relevant information, and it reduces working memory demands, thus facilitating integration (Symons & Pressley, 1993). The findings of this study suggest that the irrelevant constructs that may be introduced by computerized tests (i.e., short-term memory and spatial ability) may not be influential when the test material is easy or

familiar to examinees. However, when the material is difficult, these additional constructs may become more dominant, affecting the cognitive processes and strategies used on the test.

An important limitation of this study is the lack of a time limit imposed on participants, because actual testing situations include time limits. If a time limit had been imposed in this study, the mode differences found on the first test would be expected to have been amplified. Since it was found that participants spent more time on the first test, regardless of mode, the lack of a time limit may have compensated for the difficulty of the test, giving participants extra time to comprehend the passage.

Throughout this study, an assumption was made that computerized tests introduce irrelevant constructs into a test. However, some of the findings of this study suggest that computerized tests may actually be a better measure of reading comprehension than paper-and-pencil tests. For example, computerized tests may prevent examinees from relying on physical aids and instead compel them to pay more attention and remember information from the passage. Furthermore, computerized tests may discourage examinees from previewing the questions, making the task more of a reading comprehension exercise than a search and matching exercise.

The findings of this study offer several areas worthy of future research. Since the most salient mode difference found in this study was in the use of physical aids, the benefits of these physical aids should be further investigated. As suggested in this study, physical aids may offer comfort to examinees who are used to being able to write on the passage and test questions, but they may not offer any additional cognitive or strategic benefits. Other types of test items should also be examined, such as mathematics and analytical items which rely heavily on the use of physical aids, and which might be expected to show greater mode effects. Future studies might also explore the relationship between test mode and item interest level by obtaining measures of participants' prior knowledge and interest level and examining the effect on the strategies used to answer the test items in the two test modes. Finally, future research should highlight the ways in which computerized tests may improve construct validity. This is especially important given that more and more important tests are being computerized, and that computerized tests may completely replace paper-and-pencil tests in the not too distant future.

References

Afflerbach, P.P. (1990). The influence of prior knowledge on expert readers' main idea construction strategies. Reading Research Quarterly, 25 (1), 31-46.

Afflerbach, P., & Johnston, P. (1984). Research methodology on the use of verbal reports in reading research. Journal of Reading Behavior, 16 (4), 307-321.

Anderson, T.H., & Armbruster, B.B. (1984). Studying. In P.D. Pearson (Ed.), Handbook of reading research: Vol. 1 (pp. 657-679). New York: Longman.

Bereiter, C., & Bird, M. (1985). Use of thinking aloud in identification and teaching of reading comprehension strategies. Cognition and Instruction, 2 (2), 131-156.

Bishop, N.S., & Frisbie, D.A. (1998, December). The effects of different test-taking conditions on reading comprehension test performance. Paper presented at the Annual Meeting of the Iowa Educational Research and Evaluation Association, Ames, IA.

Bordens, K.S., & Abbott, B.B. (1991). Research design and methods: A process approach (2nd ed.). Mountain View, CA: Mayfield Publishing Company.

Chandler, P., & Sweller, J. (1991). Cognitive load theory and the format of instruction. Cognition and Instruction, 8 (4), 293-332.

Cohen, A.D. (1986). Mentalistic measures in reading strategy research: Some recent findings. English for Specific Purposes, 5 (2), 131-145.

Daniel, D.B. (1983). The construct of legibility in the reading environment of a microcomputer. (ERIC Document Reproduction Service No. ED 255 908)

Dillon, A. (1992). Reading from paper versus screens: A critical review of the empirical literature. Ergonomics, 35 (10), 1297-1326.

Educational Testing Service (1997). PowerPrep software for the GRE General Test (Version 1.1) [Computer software]. Princeton, NJ: Author.

Educational Testing Service (1996). GRE practicing to take the general test big book. Princeton, NJ: Author.

Ericsson, K.A., & Simon, H.A. (1993). Protocol analysis: Verbal reports as data. (Rev. ed.). Cambridge, MA: The MIT Press.

Farr, R., Pritchard, R., & Smitten, B. (1990). A description of what happens when an examinee takes a multiple-choice reading comprehension test. Journal of Educational Measurement, 27 (3), 209-226.

Guthrie, J.T. (1988). Locating information in documents: Examination of a cognitive model. Reading Research Quarterly, 23 (2), 178-199.

Haas, C., & Hayes, J.R. (1986). What did I just say? Reading problems in writing with the machine. Research in the Teaching of English, 20 (1), 22-35.

Heppner, F.H., Anderson, J.G.T., Farstrup, A.E., & Weiderman, N.H. (1985). Reading performance on a standardized test is better from print than from computer display. Journal of Reading, 28 (4), 321-325.

Johnston, P.H. (1984). Assessment in reading. In P.D. Pearson (Ed.), Handbook of reading research: Vol. 1 (pp. 147-182). New York, NY: Longman.

Keppel, G. (1991). Design and analysis: A researcher's handbook (3rd ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

Mazzeo, J., & Harvey, A.L. (1988). The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature (College Board Rep. No. 88-8, ETS RR No. 88-21). Princeton, NJ: Educational Testing Service.

Mead, A.D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. Psychological Bulletin, 114 (3), 449-458.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), Educational Measurement (3rd ed, pp.13-103). New York: Macmillan.

Pearson, P.D., Roehler, L.R., Dole, J.A., & Duffy, G.G. (1992). Developing expertise in reading comprehension. In S.J. Samuels & A.E. Farstrup (Eds.), What research has to say about reading comprehension (pp. 145-199). Newark, DE: International Reading Association.

Pressley, M., & Afflerbach, P. (1995). Verbal protocols of reading: The nature of constructively responsive reading. Hillsdale, NJ: Lawrence Erlbaum Associates.

Samuels, S.J., & Kamil, M.L. (1984). Models of the reading process. In P.D. Pearson (Ed.), Handbook of reading research: Vol. 1 (pp. 185-224). New York, NY: Longman.

Sternberg, R.J. (1991). Are we reading too much into reading comprehension tests? Journal of Reading, 34 (7), 540-545.

Sweller, J., & Chandler, P. (1994). Why some material is difficult to learn. Cognition and Instruction, 12 (3), 185-233.

Sweller, J., & Chandler, P. (1991). Evidence for cognitive load theory. Cognition and Instruction, 8 (4), 351-362.

Symons, S., & Pressley, M. (1993). Prior knowledge affects text search success and extraction of information. Reading Research Quarterly, 28 (3), 251-259.

Vispoel, W.P., Hendrickson, A.B., Bleiler, T., Widiatmo, H., Sharairi, S., & Ihrig, D. (1999, April). Limiting answer review and change on computerized adaptive vocabulary tests: Psychometric and attitudinal results. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.

APPENDIX
Coding Categories at the Utterance Level

Code	Description	Examples
1. Reading A. Directions B. Sentence C. Question Stem D. Answer Choice	Verbatim or close to verbatim reading of directions, passage, question stems, and answer choices.	"Ragtime is a musical form that synthesizes folk melodies and musical techniques into a brief, quadrille-like structure, designed to be played, exactly as written, on the piano." "According to the passage, each of the following is a characteristic of ragtime compositions that follow the classic ragtime formula except..."
2. Paraphrasing	Summarizing or rephrasing part(s) of the passage.	"So they're not concerned with the development of themes."
3. Skimming	Reading a few words per sentence within one paragraph or across two or more paragraphs.	"The classic formula...da da da..bright memorable strain or theme followed by a similar...lyrical strain..."
4. Identifying Important Information	Indicating that a word, phrase, or idea in the passage, question stem, or answer choice is important; or using concepts or words that are repeated to decide what is important.	"All right, so I'm underlining syncopated counterpart." "That sounds important cause he introduced that - the multireel picture." "It's talking a lot about the camera."
5. Making Inferences	Using background knowledge to fill in deleted information, elaborate on text, or draw conclusions. Using inference to generate an answer to a question before reading or evaluating the given answer choices.	"It sounded like he was, he was, the author was praising...um..this guy Griffith.." "Um..it's-cause it's played like a machine it must be well-defined melodically.." "Well, if one- four reels are one hour, then I assume that one reel is 15 minutes or less."
6. Understanding A. Positive B. Negative	Overtly stating understanding or lack of understanding of a word, sentence, concept, question stem, or answer choice; or asking questions that indicate lack of understanding.	"I just realized it's kind of like ...what the ragtime is just thinking about what that is." "I'm not really.. registering, so I have to read over a little bit." "Uh..I really don't know what they mean by composition."

Code	Description	Examples
7. Recalling	Stating that they remember (or do not remember) an idea or the location of a word or idea in the passage or questions;	"Cause I remember them saying, um... uh..about the..recording characteristics..of the..of the music..and how it sounded mechanical."
A. Content		
B. Location	(While reading passage) stating an idea from an earlier part of the passage without returning to that portion of the passage; or	"I just like, you know, assumed that, since I remembered these names were in this top paragraph, so I just scrolled up immediately, because, they're just right there."
C. Content-negative		
D. Location-negative	(While answering questions) stating an idea from the passage without returning to the passage.	"I don't know if it mentioned I don't remember reading the time.." "I don't remember..exactly where it was mentioned so I'm skimming through the paragraphs to see..."
8. Integrating Text	Noting different parts of the passage (e.g., introduction, examples, final point); noting coherence or lack of coherence between different parts of the passage; relating information currently read to information read previously; using knowledge of paragraph structure to understand passage; and attempting to get the larger meaning of the entire passage or parts of the passage.	"Um...well in the first, these two paragraphs it kind of talks about...like what it is.....[writes notes in margin], and not until the last two paragraphs do they really distinguish it from jazz.....[writes notes]. "It just seems to go into a lot of detail about..the style of..ragtime in itself and not..necessarily always..um..comparing it or contrasting it to anything else.."
9. Executive Processes		
A. Initiating Search	Stating that they are going to reference or search the passage for information to answer a question. The target of the search (word or idea), if evident, is also coded.	"I don't really know but let me see if it says anything about bass line...in here....."
B. Monitoring Search	Stating progress in relation to a search that has been initiated.	"So I'm just gonna look down...and I'm scanning...I'm still scanning..."
C. Monitoring Progress	Stating perceptions of their knowledge or ability with regard to answering the test questions, their readiness to move to the next question, and whether strategies are effective.	"I think I'm doing really bad [laughs]." "OK, I'm spending too long on this." "Wait a minute let me make sure I got that.."
D. Monitoring Location	Stating perceptions of their location within the passage or the test questions.	"Where am I" "Um..it's difficult finding..the right spot when you can't see it all at once.."
E. Stating Strategy	Stating an action or strategy related to reading the passage or answering the test questions.	"Um..actually, I'm just gonna skip ahead to the questions." "So..I'm gonna go back and check that fourth answer out but I'll wait until I finish this one.."

Code	Description	Examples
10. Using Physical Aids A. Underlining/ Marking Text B. Writing C. Crossing Out Answer Choices D. Marking Answer Choices	Using physical aids while answering the test questions	"And I'm gonna underline varying speed and rhythm." "I'm writing down A through E on my paper [writes on scratch paper].." "OK...D is out, cause it says it's not concerned with the development of musical themes [crosses out answer choice]." "Well, I think that could be it [marks choice], but we'll go on"
11. Evaluating		
A. Answer choices	Rejecting an answer choice, or considering the choice as a possible answer.	"A is a possible answer - cause I know it was talking about the contrast between that and jazz." "No, the article doesn't discuss commercial success [F] at all."
B. Search	Indicating whether a search for information to answer a question is successful or unsuccessful.	"Here we are here's something that talks about mechanical." "I don't see that anywhere, unless its at the very beginning.."
C. Relation (matches)	Indicating that information in the passage is relevant to an answer choice (i.e., the information "matches" the answer choice, or provides evidence for or against the answer choice).	"Cause it talks about the performing style and..here..it says it is not precision limited to the style of performance.." "Now it says it has become standard ever since but it doesn't say that he..necessarily introduced them to American.."
12. Using prior work	While answering the questions, returning to information in the passage that was previously identified as important information.	"Now let me read those-I'm gonna read the few paragraphs that, the little sentences I bracketed, cause they probably say the main points and they'll tell me which answer's right."
13. Selecting Answer	Choosing an answer choice.	"OK..the answer would be to define ragtime music as an art form..."
14. Navigation	Moving from one question to another.	N/A
15. Miscellaneous	Processes that cannot be coded, i.e., are ambiguous.	N/A

BEST COPY AVAILABLE